Computing a Nearest Symmetric Positive Semidefinite Matrix

Nicholas J. Higham
Department of Mathematics
University of Manchester
Manchester M13 9PL, England

Submitted by G. W. Stewart

ABSTRACT

The nearest symmetric positive semidefinite matrix in the Frobenius norm to an arbitrary real matrix A is shown to be $(B + H)/2$, where H is the symmetric polar factor of $B = (A + A^T)/2$. In the 2-norm a nearest symmetric positive semidefinite matrix, and its distance $\delta_2(A)$ from A, are given by a computationally challenging formula due to Halmos. We show how the bisection method can be applied to this formula to compute upper and lower bounds for $\delta_2(A)$ differing by no more than a given amount. A key ingredient is a stable and efficient test for positive definiteness, based on an attempted Choleski decomposition. For accurate computation of $\delta_2(A)$ we formulate the problem as one of zero finding and apply a hybrid Newton-bisection algorithm. Some numerical difficulties are discussed and illustrated by example.

1. INTRODUCTION

Symmetric positive definiteness is arguably one of the highest mathematical accolades to which a matrix can aspire. For symmetry confers important advantages and simplifications in the eigenproblem, and positive definiteness permits economy and numerical stability in the solution of linear systems. Thus it is pleasing that in a wide variety of physical problems the application of minimum principles gives rise to symmetric positive definite linear systems [19]. But in some applications a matrix that is expected to be symmetric positive (semi)definite fails to be so, and it is required to approximate it by a matrix that is. A well-known example occurs in detecting and dealing with an indefinite Hessian in optimization [10]. In another example, mentioned in [16], it is required to compensate for the errors of measurement, computation,

52 Vanderbilt Ave., New York, NY 10017 0024-3795/88/$3.50

or even typing, which could vitiate the symmetry or definiteness of a matrix. Some other problems with matrix definiteness constraints are described in [8].

This work is concerned with the computation of nearest symmetric positive semidefinite matrices to an arbitrary matrix, in the 2-norm and in the Frobenius norm. (We will use the abbreviation psd for "symmetric positive semidefinite.") Our work adds to the recent literature on matrix nearness problems [3, 5, 13, 14, 20].

In Section 2 we show that in the Frobenius norm there is a unique nearest psd matrix to A, which can be expressed in terms of the polar decomposition of the symmetric part of A. This nearest matrix is easily computed using a single spectral decomposition.

In the 2-norm there are, in general, many nearest psd matrices to A. Halmos [12] obtains one such matrix in terms of a convenient, though computationally challenging, formula for its distance from A. We use Halmos's formula to derive methods for estimating and computing this matrix and its 2-norm distance from A.

Halmos's result is stated and explored in Section 3. In Section 4 we develop a bisection algorithm, similar to that in [3], for computing upper and lower bounds on the 2-norm distance to the nearest psd matrix. To implement the algorithm we need an efficient and numerically stable method for testing whether a given symmetric matrix is positive definite. We show in Section 5 that an attempted Cholesky factorization has the required properties, and we investigate a "reverse pivoting" implementation that attempts to reduce the cost.

In Section 6 we consider the use of a hybrid bisection Newton iteration for computing the distance to high accuracy. Section 7 presents some illustrative numerical examples.

Our notation is as follows. All matrices are assumed real (there is no difficulty in extending all the results and algorithms to the complex case). We use the 2-norm,

$$\|A\|_2 = \rho(A^TA)^{1/2},$$

where the spectral radius $\rho(B) = \max\{|\lambda| : \det(B - \lambda I) = 0\}$, and the Frobenius norm

$$\|A\|_F = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}.$$

Recall that a symmetric matrix A is positive definite if its eigenvalues are positive, and positive semidefinite, which we will denote by $A \geq 0$, if its

eigenvalues are nonnegative. Following Halmos [12], the distance in the norm $\|\cdot\|$ from an arbitrary A to the set of psd matrices is denoted by

$$\delta(A) = \min_{X = X^T \geq 0} \|A - X\|,$$

and any psd X satisfying $\|A - X\| = \delta(A)$ is termed a *positive approximant* of A in the given norm. For a psd A, $A^{1/2}$ denotes the unique psd matrix X satisfying $X^2 = A$.

It is important when solving a numerical problem to understand the limitations imposed by finite precision arithmetic. Whatever method is used to compute the distance $\delta(A)$, we cannot expect better than to obtain the true distance for a slightly perturbed matrix, that is, a value $\hat{\delta} = \delta(A + E)$ where $\|E\| \leq \epsilon\|A\|$, ϵ being a small multiple of the working precision. Then

$$|\hat{\delta} - \delta(A)| \leq \|E\| \leq \epsilon\|A\|, \tag{1.1}$$

which gives an upper bound $\epsilon\|A\|$ on the absolute error. Writing (1.1) in terms of the scale independent quantity $\delta(A)/\|A\| \leq 1$, we have, for $\delta(A) \neq 0$,

$$\frac{\left| \dfrac{\hat{\delta}}{\|A\|} - \dfrac{\delta(A)}{\|A\|} \right|}{\dfrac{\delta(A)}{\|A\|}} \leq \epsilon \left(\frac{\delta(A)}{\|A\|} \right)^{-1}.$$

Hence the bound on the relative accuracy with which we can compute $\delta(A)/\|A\|$ depends on the size of the quantity itself; generally, the smaller it is, the fewer significant figures we can obtain. Interestingly, the same conclusion applies to any relative distance function, such as the relative distance to the nearest singular matrix (which for the 2-norm is the reciprocal of $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$).

2. THE FROBENIUS NORM POSITIVE APPROXIMANT

The following result gives the solution to the problem of positive approximation in the Frobenius norm. The result appears to be new.

THEOREM 2.1. *Let $A \in \mathbb{R}^{n \times n}$, and let $B = (A + A^T)/2$ and $C = (A - A^T)/2$ be the symmetric and skew-symmetric parts of A respectively.*

Let $B = UH$ be a polar decomposition ($U^TU = I$, $H = H^T \geqslant 0$). Then $X_F = (B + H)/2$ is the unique positive approximant of A in the Frobenius norm, and

$$\delta_F(A)^2 = \sum_{\lambda_i(B) < 0} \lambda_i(B)^2 + \|C\|_F^2.$$

Proof. Let X be psd. From the fact that $\|S + K\|_F^2 = \|S\|_F^2 + \|K\|_F^2$ if $S = S^T$ and $K = -K^T$, we have

$$\|A - X\|_F^2 = \|B - X\|_F^2 + \|C\|_F^2,$$

and so the problem reduces to that of approximating B. Let $B = Z\Lambda Z^T$ [$Z^TZ = I$, $\Lambda = \operatorname{diag}(\lambda_i)$] be a spectral decomposition, and let $Y = Z^TXZ$. Then

$$\begin{aligned}\|B - X\|_F^2 &= \|\Lambda - Y\|_F^2 \\ &= \sum_{i \neq j} y_{ij}^2 + \sum_i (\lambda_i - y_{ii})^2 \\ &\geqslant \sum_{\lambda_i < 0} (\lambda_i - y_{ii})^2 \geqslant \sum_{\lambda_i < 0} \lambda_i^2,\end{aligned}$$

since $y_{ii} \geqslant 0$ because Y is psd. This lower bound is attained, uniquely, for the matrix $Y = \operatorname{diag}(d_i)$, where

$$d_i = \begin{cases} \lambda_i, & \lambda_i \geqslant 0, \\ 0, & \lambda_i < 0, \end{cases} \tag{2.1}$$

that is,

$$X_F = Z \operatorname{diag}(d_i) Z^T. \tag{2.2}$$

The representation $X_F = (B + H)/2$ follows, since $H = Z \operatorname{diag}(|\lambda_i|) Z^T$. ∎

Computation of the positive approximant $X_F = (B + H)/2$ is straightforward. The preferred approach is to compute a spectral decomposition of B and to form X_F according to (2.1) and (2.2). An alternative, when B is nonsingular, is to compute H using the iterative algorithm of [13], which is based on matrix inversions.

3. A 2-NORM POSITIVE APPROXIMANT

Halmos [12] proves the following result, in the more general context of linear operators on a Hilbert space.

THEOREM 3.1. *For* $A \in \mathbb{R}^{n\times n}$,

$$\delta_2(A) = \min\{ r \geq 0 : r^2 I + C^2 \geq 0 \text{ and } B + (r^2 I + C^2)^{1/2} \geq 0 \}, \quad (3.1)$$

where $B = (A + A^T)/2$ *and* $C = (A - A^T)/2$ *are the symmetric and skew-symmetric parts of* A *respectively. The matrix*

$$P = B + \left[\delta_2(A)^2 I + C^2\right]^{1/2} \quad (3.2)$$

is a positive approximant of A.

The importance of Halmos's result is that it replaces the problem of minimizing over the set of $n \times n$ psd matrices by the much simpler problem of minimizing over the nonnegative scalars.

It should be stressed that a 2-norm positive approximant of A is not in general unique. To see this, note that $B = (A + A^T)/2$ is a nearest symmetric matrix to A in the 2-norm [7]. If B is psd, then clearly it must also be a positive approximant for A; but if C^2 is not a multiple of the identity, then B differs from the positive approximant given by (3.2). For results on the uniqueness of positive approximants see [1, 2, 12].

We now examine some properties of the matrix

$$G(r) = B + (r^2 I + C^2)^{1/2}$$

which occurs in (3.1). The eigenvalues of $r^2 I + C^2$ are $r^2 - |\lambda_i(C)|^2$, since $\lambda_i(C) \in i\mathbb{R}$ because of C's skew-symmetry. Thus the condition $r^2 I + C^2 \geq 0$ in (3.1) is equivalent to $r \geq \rho(C)$. The next two results, combined with Theorem 3.1, form the basis for the numerical methods to be developed in Sections 4 and 6.

LEMMA 3.2. $\lambda_{\min}(G(r))$ *is a strictly monotone function of* r $(r \geq \rho(C))$.

Proof. Let $r \geq \rho(C)$ and $\delta r > 0$. Then $(r + \delta r)^2 I + C^2$ is positive definite, and hence so is

$$D = \left[(r + \delta r)^2 I + C^2\right]^{1/2} + (r^2 I + C^2)^{1/2}.$$

We have

$$G(r+\delta r)-G(r)=\left[(r+\delta r)^2 I+C^2\right]^{1/2}-(r^2 I+C^2)^{1/2}$$

$$=D^{-1}\left[(r+\delta r)^2-r^2\right]I=(2r\delta r+\delta r^2)D^{-1},$$

using the fact that the two matrices whose sum is D commute. Thus $G(r+\delta r)-G(r)$ is positive definite, which implies that $\lambda_{\min}(G(r+\delta r))-\lambda_{\min}(G(r))>0$, as required. ∎

COROLLARY 3.3. *Suppose $G(\rho(C))$ is not positive semidefinite. Then there is a unique $r^*>\rho(C)$ for which $G(r^*)$ is positive semidefinite and singular. For $r>r^*$, $G(r)$ is positive definite, and for $r<r^*$, $G(r)$ has at least one negative eigenvalue.*

The next result shows that from an estimate r for $r^*=\delta_2(A)$ we obtain a matrix $G(r)$ whose distance from A is correspondingly close to the minimum, $\delta_2(A)$.

LEMMA 3.4. $\|A-G(r)\|_2=r$ $(r\geqslant\rho(C))$.

Proof.

$$\|A-G(r)\|_2=\left\|C-(r^2I+C^2)^{1/2}\right\|_2$$

$$=\max_{i\mu\in\lambda(C)}\left|i\mu-(r^2-\mu^2)^{1/2}\right|$$

$$=r.$$ ∎

Finally we consider the role of the Frobenius norm positive approximant in approximation in the 2-norm. When A is normal (that is $AA^T=A^TA$), $X_F=(B+H)/2$ is actually a 2-norm positive approximant of A. This is shown by Halmos [12]. Note, however, that X_F is generally different from P in (3.2). For arbitrary A we can show that X_F is an approximate minimizer of the 2-norm distance $\|A-X\|_2$.

LEMMA 3.5. $\delta_2(A)\leqslant\|A-X_F\|_2\leqslant 2\delta_2(A)$.

Proof. The lower bound is trivial. For the upper bound,

$$\|A - X_F\|_2 = \|C + \tfrac{1}{2}(B - H)\|_2 \leq \|C\|_2 + \tfrac{1}{2}\|B - H\|_2$$

$$= \rho(C) + \max\{0, -\lambda_{\min}(B)\},$$

since $B - H = Z\,\mathrm{diag}(\lambda_i - |\lambda_i|)Z^T$, $Z^TZ = I$ (see the proof of Theorem 2.1). We have $\rho(C) \leq \delta_2(A)$. Also

$$0 \leq \lambda_{\min}\left(B + \left[\delta_2(A)^2 I + C^2\right]^{1/2}\right) \leq \lambda_{\min}(B) + \lambda_{\max}\left(\left[\delta_2(A)^2 I + C^2\right]^{1/2}\right)$$

$$\leq \lambda_{\min}(B) + \delta_2(A),$$

since C^2 has negative eigenvalues; thus

$$\max\{0, -\lambda_{\min}(B)\} \leq \delta_2(A),$$

as required. ∎

4. A BISECTION METHOD FOR ESTIMATING $\delta_2(A)$

Except in special cases (such as when A is normal) there is no direct way to compute $\delta_2(A)$ from (3.1), because the eigenvalues of $G(r) = B + (r^2 I + C^2)^{1/2}$ are nonlinear functions of r. Therefore we turn to iterative methods. We suppose in this section that $\delta_2(A)$ is to be estimated to modest accuracy.

Suppose $G(\rho(C))$ is not positive semidefinite [if it is, then $\delta_2(A) = \rho(C)$]. Corollary 3.3 suggests the following bisection approach. Find an interval $[\alpha, \beta]$ containing $r^* = \delta_2(A)$. If $G(\gamma)$ is positive definite, where $\gamma = (\alpha + \beta)/2$, accept the interval $[\alpha, \gamma]$ containing r^*; otherwise accept $[\gamma, \beta]$. Repeat the process until the desired accuracy is obtained.

A suitable convergence test is (since $0 \leq \alpha < \beta$)

$$\frac{\beta - \alpha}{2} \leq \max\{f\alpha, \mathrm{tol}\}, \tag{4.1}$$

where $f < 1$ is a relative error tolerance, and tol is an absolute error tolerance which takes effect when α is zero or very small. A suitable choice for tol is $\mu\|A\|_F$, where μ should be no smaller than the machine precision [cf. (1.1)].

The cost of forming $G(\gamma)$ at each stage can be reduced considerably by making use of a spectral decomposition of C^2. If

$$C^2 = Z\Lambda Z^T, \qquad Z^TZ = I, \quad \Lambda = \operatorname{diag}(\lambda_i),$$

then

$$\begin{aligned} G(r) &= B + Z(r^2I + \Lambda)^{1/2}Z^T \\ &= Z\left[Z^TBZ + (r^2I + \Lambda)^{1/2}\right]Z^T, \end{aligned}$$

and it suffices to form and test for positive definiteness the matrix $\tilde{B} + (r^2I + \Lambda)^{1/2}$, where $\tilde{B} = Z^TBZ$, and where the square root is trivial to compute.

An initial interval containing r^* is given by the bounds

$$\max\left\{\rho(C), \max_{\tilde{b}_{ii}<0} \sqrt{\tilde{b}_{ii}^2 - \lambda_i(C^2)}, M\right\} \leqslant r^* \leqslant \rho(C) + M, \tag{4.2}$$

where

$$M = \max\{0, -\lambda_{\min}(B)\}.$$

The upper bound is from the proof of Lemma 3.5, and the second lower bound follows from the fact that a psd matrix has nonnegative diagonal elements. These bounds differ by no more than a factor 2, and they are exact when A is symmetric (since $C = 0$). We mention that computation of $\lambda_{\min}(B)$ could be avoided by using the alternative, but potentially much weaker, upper bound $\|A\|_F$.

We obtain the following algorithm.

ALGORITHM EST.

Input: $A \in \mathbf{R}^{n\times n}$, a relative error tolerance $f < 1$, and an absolute error tolerance tol.

Output: $\alpha, \beta \geqslant 0$ such that

$$\alpha \leqslant \delta_2(A) \leqslant \beta \leqslant \alpha + 2\max\{f\alpha, \text{tol}\},$$

and a psd matrix X such that $\|A - X\|_2 = \beta$.

1. $B := (A + A^T)/2$; $C := (A - A^T)/2$.
2. $C^2 = Z\Lambda Z^T$ (spectral decomposition).

3. $B := Z^T B Z$.
4. Form an interval $[\alpha, \beta]$ bracketing $\delta_2(A)$ using (4.2).
5. If $B + (\alpha^2 I + \Lambda)^{1/2}$ is psd set $\beta := \alpha$ and goto step 7.
6. While $(\beta - \alpha)/2 > \max\{f\alpha, \text{tol}\}$
   $r := (\alpha + \beta)/2$,
   $G := B + (r^2 I + \Lambda)^{1/2}$,
   if G is psd then $\beta = r$ else $\alpha = r$.
7. $X := Z[B + (\beta^2 I + \Lambda)^{1/2}]Z^T$.

The test in step 5 identifies the cases where the lower bound is exact and hence sectioning is unnecessary.

Algorithm EST works entirely in real arithmetic. In step 2 one can either form C^2 explicitly and apply a symmetric eigensolver, or compute a real Schur decomposition $C = ZDZ^T$ (preferably using a routine that takes advantage of skew-symmetry) and take $\Lambda = D^2$. The former approach is numerically stable, since $\|C\|_2^2 = \|C^2\|_2$.

A remaining question is how to test the definiteness of G in the iteration of step 6. This is considered in the next section.

5. TESTING FOR POSITIVE DEFINITENESS

In Algorithm EST we require a method for determining whether a given symmetric matrix G is positive semidefinite. The method should be efficient and numerically stable. By the latter we mean that the answer, "yes" or "no," should be the correct answer for a nearby symmetric matrix

$$\tilde{G} = G + E, \qquad \|E\|_2 \leq \epsilon \|G\|_2,$$

where ϵ is a small multiple of the machine precision. Since with an arbitrarily small perturbation a semidefinite matrix can become definite, it follows that in finite precision arithmetic testing for definiteness is equivalent to testing for semidefiniteness. We will adopt the viewpoint of testing for strict definiteness.

One approach is to compute the eigenvalues of G, using the symmetric QR algorithm [11, p. 281], and to check if the smallest computed eigenvalue is positive. This approach is certainly numerically stable, since the computed eigenvalues are those of a nearby matrix [11, p. 282]. The cost is about $2n^3/3$ flops.

A less expensive approach is to attempt to compute a Cholesky decomposition of G, declaring the matrix positive definite if the process succeeds,

that is, if no zero or negative pivots are encountered, and not positive definite otherwise. The cost is at most $n^3/6$ flops, and depends on the number of successful elimination stages. This Cholesky approach is related to techniques to be found in [17, p. 46], where LDL^T factorizations are used to compute the inertia of a symmetric matrix, and in [4], where the inertia of a symmetric Toeplitz matrix is computed using the Levinson-Durbin algorithm.

The stability, or otherwise, of the Cholesky definiteness test does not seem obvious. Indeed, at first sight it is not clear why instability cannot contrive to allow Cholesky factorization of a "safely" indefinite matrix to succeed. To investigate the stability we turn to the classical error analysis for the Cholesky decomposition.

THEOREM 5.1 [15, 21]. *Let $G \in \mathbb{R}^{n\times n}$ be a positive definite matrix of floating point numbers. Then there are small constants c_n and d_n such that*

(a) *the Cholesky algorithm runs to completion if $\kappa_2(G)c_n u < 1$; and*
(b) *the computed Cholesky factor $\hat{L}$ satisfies*

$$\hat{L}\hat{L}^T = G + E, \qquad \|E\|_2 \leq d_n u \|G\|_2. \qquad \blacksquare \quad (5.1)$$

Theorem 5.1 immediately yields one half of the required stability result. For if Cholesky factorization of G breaks down, then G cannot be positive definite with $\kappa_2(G)c_n u < 1$ as this would contradict part (a). Thus either G is not positive definite or it is positive definite with $\kappa_2(G)c_n u > 1$. In the latter case G is within 2-norm distance $c_n u\|G\|_2$ of a singular, symmetric matrix, and the answer "not positive definite" is therefore the correct one for a matrix close to G.

Considering now the "positive definite" answer, we note that Theorem 5.1 apparently is not applicable to general symmetric matrices G. However, examination of the proofs in [15, 21] reveals that the backward error result (5.1) holds without any assumption on the definiteness of G—it requires only that the algorithm run to completion. It follows from (5.1) that if Cholesky factorization of a symmetric matrix G succeeds, then G is near to a positive definite matrix, namely $\hat{L}\hat{L}^T$.

We conclude that the Cholesky positive definiteness test is numerically stable.

A modification to the Cholesky test which can reduce the cost without compromising the stability was suggested by C. F. Van Loan (private communication). The idea is to use a "reverse pivoting" version of the Cholesky algorithm, in which at each stage one chooses as pivot the *smallest* element among the remaining diagonal elements, with the aim of inducing

early breakdown of the process when A is not positive definite. To investigate the effectiveness of reverse pivoting we considered the SAXPY implementation of the Choleski algorithm, as used in LINPACK's SCHDC [6], in which at the kth stage a rank 1 matrix is subtracted to zero out the kth row and column, at a cost of $(n-k)^2/2$ flops. It is easy to construct examples in which either of the reverse pivoting and no-pivoting SAXPY versions terminates at an earlier stage than the other. We ran some numerical tests using a modified version of SCHDC, both within Algorithm EST (see Section 7) and on random symmetric matrices with various spectra, with $n \leq 25$. With only one exception, the reverse pivoting version terminated at the same, or an earlier, stage than the no-pivoting version. However, considering that approximately half the work of a complete factorization is done after $[n/5]$ stages, the computational savings brought by the reverse pivoting implementation were relatively small. In fact, we recommend the use of the no-pivoting SDOT Choleski algorithm used in LINPACK's SPOFA [6], combined with an initial scan for nonpositive diagonal elements. Note that in this implementation the kth stage costs $k^2/2$ flops, so for matrices that are not positive definite, less work will be required than with the no-pivoting SAXPY version.

6. ACCURATE DETERMINATION OF $\delta_2(A)$

If $\delta_2(A)$ is to be computed to many significant figures, then it is desirable to use a method that converges more rapidly than Algorithm EST. To obtain such a method we assume that $\delta_2(A) > \rho(C)$ and we use Theorem 3.1, Lemma 3.2, and Corollary 3.3 to reformulate the problem of computing $\delta_2(A)$ as that of finding the unique zero of the function

$$f(\tau) = \lambda_{\min}(G(\tau)), \tag{6.1}$$

where

$$G(\tau) = B + (\tau^2 I + C^2)^{1/2}.$$

Provided $\lambda_{\min}(G(\tau))$ is a simple eigenvalue, then f is differentiable at $\tau > \rho(C)$, and, using standard analysis [11, p. 202],

$$\begin{aligned} \dot{f}(\tau) &= x(\tau)^T \dot{G}(\tau) x(\tau) \\ &= \tau x(\tau)^T (\tau^2 I + C^2)^{-1/2} x(\tau), \end{aligned} \tag{6.2}$$

where $x(r)$ is an eigenvector corresponding to $\lambda_{\min}(G(r))$, normalized so that $\|x(r)\|_2 = 1$. We can therefore apply Newton's method, suitably constrained to the interval $[\rho(C), \infty)$ on which f is defined. We will use a "fail-safe" hybrid Newton-bisection algorithm, in which a bisection step is taken if the Newton step either would leave the current bracket or would not produce a sufficient reduction in the size of the bracket.

There are several details concerning the implementation.

(1) If A is normal, then X_F is a 2-norm positive approximant and both X_F and $\delta_2(A)$ may be computed from a single spectral decomposition, that of A (see Sections 2 and 3). This will be less expensive than the Newton-bisection approach, and so normal matrices should be treated as a special case.

(2) An initial bracket can be obtained from (4.2). A spectral decomposition of C should be used as in Algorithm EST to reduce the cost of evaluating f and $\dot{f}$.

(3) If $\lambda_{\min}(G(r))$ is a repeated eigenvalue, then f is not differentiable at r; the expression in (6.2) exists, but it is not uniquely defined. Finding the zero of f in (6.1) can be posed as a nonlinear inverse eigenvalue problem: Minimize r subject to $\lambda_i(G(r)) \geq 0$ for all i. The behavior of Newton's method for solving *linear* inverse eigenvalue problems is investigated in [9]. It is shown that local quadratic convergence is generally obtained, both in theory and in practice, even when there are multiple eigenvalues at the solution. It seems reasonable to expect this behavior to hold also for our nonlinear problem. In any case, the hybrid Newton-bisection algorithm is guaranteed to converge, by construction, albeit only linearly in the worst case. A pleasing result in these circumstances is that the positive approximant $P = G(\delta_2(A))$ has the minimum number of zero eigenvalues over all positive approximants of A. This follows from Theorem 4.2 in [1], which states that $P - X \geq 0$ for any 2-norm positive approximant X of A.

(4) Depending on the relative separation of r, $\rho(C)$, and $\delta_2(A)$, the Newton step may leave the range $[\rho(C), \infty)$ on which $f(r)$ is defined, causing a bisection step to be taken. By considering the $n = 1$ case, $f(r) = b + \sqrt{r^2 - c^2}$ $(r \geq |c|)$, which represents the upper half of a parabola, it is easy to see that more steps may be required for convergence when $\delta_2(A) \approx \rho(C)$.

(5) Since we are using a spectral decomposition of C^2, we work with $G(r)$ in the form $G(r) = \tilde{B} + [r^2 I - \operatorname{diag}(\mu_i^2)]^{1/2}$ (see Section 4). For $r \approx r^*$, when forming

$$g_{ii}(r) = \tilde{b}_{ii} + \sqrt{r^2 - \mu_i^2}, \tag{6.3}$$

loss of significant figures will occur in the subtraction when $r^* \approx \rho(C)$, and

may occur in the addition: for example when $\tilde{B}$ is diagonal, so that [if $G(r^*)$ is not positive definite] $\lambda_{\min}(G(r^*)) = g_{ii}(r^*) = 0$ (some i). The accuracy of the computed function and derivative values may be adversely affected (see Example 5 in the next section). Unfortunately there does not seem to be any way to avoid these losses of significant figures, but at least they are easily detected.

7. NUMERICAL EXAMPLES

We have implemented Algorithm EST, and the zero finding approach of section 6, in FORTRAN 77 on a CDC Cyber computer with machine precision $u \approx 3.55 \times 10^{-15}$. We used the NAG Library spectral decomposition routine F02ABF for all eigencomputations. For zero finding by hybrid Newton-bisection we used subroutine RTSAFE from [18], specifying the stringent absolute error tolerance $u\|A\|_F$ [see (1.1)]. An initial bracket was obtained from (4.2). In Algorithm EST we took $f = 5 \times 10^{-4}$, and we tested for positive definiteness using a reverse pivoting modification of LINPACK's SCHDC [6] (see Section 5).

For each of the five examples presented we report in Table 1 the number of steps taken by Algorithm EST and by RTSAFE. We also show the sequence of steps taken by RTSAFE, with "N" for Newton and "B" for bisection, and the number of successful stages for each attempted Choleski decomposition in Algorithm EST, specified as a sequence of integers. Note that the variation in the number of steps taken by Algorithm EST reflects variation in the ratio of the bounds in (4.2).

TABLE 1

Example	n	Algorithm	Steps	Sequence (see text)
1	3	EST	10	(3, 2, 3, 3, 3, 2, 3, 2, 2, 3)
		RTSAFE	5	(NNNNN)
2	5	EST	9	(5, 5, 5, 5, 5, 3, 3, 4, 5)
		RTSAFE	10	(BBBBNNNNNN)
3	4	EST	9	(4, 4, 3, 4, 4, 4, 3, 4, 3)
		RTSAFE	7	(NNNNNNN)
4	10	EST	10	(10, 1, 1, 10, 1, 10, 1, 10, 10, 10)
		RTSAFE	5	(NNNNN)
5	5	EST	3	(4, 4, 4)
		RTSAFE	22	(NNBBNB...NBBNB)

EXAMPLE 1.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Halmos [12] states that $\delta_2(A) = \frac{1}{2}(1+\sqrt{5})^{1/2} \approx 0.8994$. The computed value of $\delta_2(A)$ was correct to 15 digits. $P = G(\delta_2(A))$ has one zero eigenvalue and two of order 1, and its upper triangle is given to four significant figures by

$$P \approx \begin{bmatrix} 0.5559 & 0.4370 & 0.5559 \\ & 0.6871 & 0.1669 \\ & & 0.7682 \end{bmatrix}.$$

For comparison, $\delta_F(A) \approx 1.2247$ and

$$X_F \approx \begin{bmatrix} 0.1768 & 0.2500 & 0.1768 \\ & 0.3536 & 0.2500 \\ & & 0.1768 \end{bmatrix}.$$

EXAMPLE 2. A is a perturbation of the 5×5 Hilbert matrix $((i+j-1)^{-1})$, obtained by setting the (4,5) element to zero. The slower convergence with the initial sequence of bisections can be explained by the closeness of $\delta_2(A) \approx 0.0632$ to $\rho(C) \approx 0.0625$ [see note (4) of Section 6].

EXAMPLE 3.

$$A = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

To five figures the computed $\delta_2(A)$ is 1.2748. $G(\delta_2(A))$ has one zero eigenvalue, the others being of order 1. The closeness of $\delta_2(A)$ to $\rho(C) \approx 1.2071$ does not unduly affect the convergence, although we did observe that many more steps were required when using a weaker initial bracket than that given by (4.2).

EXAMPLE 4. Here $A = 2ee^T - I + D$, of order 10, where $e = [1,1,\ldots,1]^T$ and D is block diagonal, composed of blocks

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

It is easy to see that $\delta_2(A) = \sqrt{2}$ and that $G(\delta_2(A)) = 2ee^T$, which has a zero eigenvalue of multiplicity 9. Despite the repeated eigenvalues, rapid convergence is obtained, with the computed $\delta_2(A)$ correct to 14 figures. For comparison, $\delta_F(A) = \sqrt{19}$ and $X_F = \frac{19}{10}ee^T$.

EXAMPLE 5. $A = \text{diag}(1, -1, -1, -1) + 0.01e_1e_4^T$, where e_i is the ith column of the identity. Here, $\delta_2(A) \approx 1.0 \gg \rho(C) = 0.005$. The very slow convergence is caused by loss of significant figures in evaluating

$$g_{44}(r) = -1 + \sqrt{r^2 - 2.5 \times 10^{-5}} = \lambda_{\min}(G(r)).$$

[$G(r)$ is diagonal.] About 5 figures are lost for $r \approx \delta(A)$.

8. CONCLUSIONS

In the Frobenius norm, the unique positive approximant of A, and the distance $\delta_F(A)$, are easily evaluated using a single spectral decomposition. Halmos's formula (3.1) for $\delta_2(A)$ poses some interesting computational problems. Algorithm EST is an efficient way to estimate $\delta_2(A)$, and a positive approximant, to low accuracy. For high accuracy computations the zero finding technique of Section 6 works well in most cases, and we have not encountered any difficulties in problems where the positive approximant $G(\delta_2(A))$ has multiple zero eigenvalues. The usefulness of the formula (3.1) for high accuracy computations is limited by its potential for loss of significant figures, which can affect the attainable accuracy and slow the convergence of an iterative algorithm. Fortunately, such loss of significance seems uncommon.

This work was begun during a visit to Stanford University in summer 1986. I thank Gene Golub for financial support and for providing a stimulating working environment. I benefited from useful discussions on this work with Ralph Byers and Michael Overton. The comments of Des Higham and Ian Gladwell helped me to improve the manuscript.

REFERENCES

1 R. Bouldin, Positive approximants, *Trans. Amer. Math. Soc.* 177:391–403 (1973).
2 R. Bouldin, Operators with a unique positive near-approximant, *Indiana Univ. Math. J.* 23:421–427 (1973).

3 R. Byers, A Bisection Method for Measuring the Distance of a Stable Matrix to the Unstable Matrices, Manuscript, Dept. of Mathematics, North Carolina State Univ., Raleigh, 1986.
4 G. Cybenko and C. F. Van Loan, Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix, *SIAM J. Sci. Statist. Comput.* 7:123–131 (1986).
5 J. W. Demmel, On Condition Numbers and the Distance to the Nearest Ill-Posed Problem, *Numer. Math.* 51:251–289 (1987).
6 J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, LINPACK *Users' Guide*, SIAM Publ., Philadelphia, 1979.
7 K. Fan and A. J. Hoffman, Some metric inequalities in the space of matrices, *Proc. Amer. Math. Soc.* 6:111–116 (1955).
8 R. Fletcher, Semi-definite matrix constraints in optimization, *SIAM J. Control Optim.* 23:493–513 (1985).
9 S. Friedland, J. Nocedal, and M. L. Overton, The Formulation and Analysis of Numerical Methods for Inverse Eigenvalue Problems, SIAM *J. Numer. Anal.* 24:634–667 (1987).
10 P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic, London, 1981.
11 G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins U.P., Baltimore, 1983.
12 P. R. Halmos, Positive approximants of operators, *Indiana Univ. Math. J.* 21:951–960 (1972).
13 N. J. Higham, Computing the polar decomposition—with applications, *SIAM J. Sci. Statist. Comput.* 7:1160–1174 (1986).
14 N. J. Higham, The symmetric procrusters problem, *BIT*, to appear.
15 J. Meinguet, Refined error analyses of Cholesky factorization, *SIAM J. Numer. Anal.* 20:1243–1250 (1983).
16 B. N. Parlett, Progress in numerical analysis, *SIAM Rev.* 20:443–456 (1978).
17 B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J., 1980.
18 W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge U.P., Cambridge, 1986.
19 G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, Mass., 1986.
20 C. F. Van Loan, How near is a stable matrix to an unstable matrix?, in *Linear Algebra and its Role in Systems Theory* (B. N. Datta, Ed.), Contemporary Math., Vol. 47, Amer. Math. Soc., 1985, pp. 465–478.
21 J. H. Wilkinson, A priori error analysis of algebraic processes, in *Proceedings of the International Congress of Mathematicians*, Moscow, 1966 (I. G. Petrovsky, Ed.), Mir, Moscow, 1968, pp. 629–640.

Received 23 February, 1987; final manuscript accepted 13 July 1987